Time Series for Forecasting Stock Market Prices Based on Sentiment Analysis of Social Media

Dakshinamoorthy Karthikeyan(✉) , Babu Aravind Sivamani ,
Pavan Kalyan Tummala , and Chamundeswari Arumugam

Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India
dakshinamoorthy17030@cse.ssn.edu.in
https://ssn.edu.in

Abstract. This paper attempts to find a relation between the public perception of a company and its stock value price. Since social media is a very powerful tool used by a lot of people to voice their opinions on the performance of a company, it is a good source of information about its public sentiment. Previous studies have shown that the overall public sentiment collected from websites like twitter do have a relation to the market price of a company over a period of time. The goal is to build on their research to improve the accuracy of predictions and determine if the public perception surrounding a company is a driving factor of its stock growth.

Keywords: Stock market · Natural Language Processing · Sentiment analysis · Social media · Time series analysis · Stock growth prediction · LSTM · Random forest

1 Introduction

Previous studies [1] on the effect of social media on the stock market have shown that the aggregate public mood towards a company over a short time span has a relation to the closing price of that company at the end of the time span. Studies have been able to utilize data collected from any one reputed social media site (Ex: Twitter, Stocktwits, Weibo, etc.) to produce a model that predicts stock market prices with 70% accuracy [2]. This paper attempts to more accurately gauge the public sentiment of a company from social media websites such as Twitter by implementing time series analysis at minute intervals to find correlations that will likely produce a better stock estimate.

Stock price of a company is determined by a large number of independent traders all over the world. Previous studies have not taken into account the reasons why an individual trader makes the decision to buy or sell. As social media has been shown to offer an insight into the mindset of people, it was realized that the posts online may be an indication of how the market at large is

inclined towards a company. The main objective of this paper is to find whether the public sentiment surrounding a company is able to determine the growth of its stock price. Here in this paper, the company Apple (NASDAQ: AAPL) was selected because it is prominent in the public spotlight and hence ideally suited for an analysis of this kind.

First the selected social media platform is queried for posts in the time period containing any of the keywords in our search term. The search term must be carefully selected to ensure that the number of off-topic posts is limited, while not missing out on any messages with important content. Then any irrelevant posts which passed through the search query are found and filtered out. Data preprocessing procedures such as the removal of non-english characters, stop words, hashtags and user mentions is carried out. Sentiment analysis is performed on the pre-processed text data and each post is classified as positive, negative or neutral corresponding to whether the market for Apple is bullish, bearish or not having any effect. Finally, the aggregate sentiment values from all collected websites will be fed into the model which would use a machine learning algorithm to produce a correlation between the media posts and the stock market price that can then be used to predict the closing market value, given the opening price and overall public sentiment.

The organization of this paper proceeds as follows. Section 2 discusses the literature survey, while Sect. 3 elaborates on the proposed methodology. Section 4 details the result and discussion, and Sect. 5 details the conclusion and future work.

2 Literature Survey

Venkata et al. [3] used Word2vec and N-gram representation of text to train a classifier model to predict the stock market movements and picked Word2vec representation due to its high accuracy in large datasets. Rakhi et al. [4] collected the sentiment data, and the stock price data to predict stock market price using a Support Vector Machine (SVM) classifier and observed that if the data size increases the accuracy obtained will also increase. Scott et al. [5] used smart user classification to filter the tweets by computing scoring weights based on number of likes, number of followers count and how often the user is correct. Further, they used Tf-Idf vectorizer for textual representation and linear regression classifier for the sentiment prediction. Zhaoxia et al. [6] used the sentiments of the news data to predict the stock market price using neural networks.

Sreelekshmy et al. [7] applied Recurrent Neural Networks(RNN), Long shortterm memory (LSTM) and Convolutional Neural Networks (CNN) - sliding window architecture for stock price prediction of Infosys, TCS and Cipla and concluded that CNN outperforms the other two models in the stock market analysis due to the irregular changes that happen in the stock market. Few works have used the previous stock market data to predict the movements of the stock market while another few used the sentiments from social media to predict the same using SVM, random forest and other machine learning algorithms. Also it is

clear that Word2Vec representation of text will be ideal for data that is fed into the neural network layers for building the classifier that predicts the trends of the stock market.

Behera R.K., et al. [16] describe a convolutional model for sentiment analysis of messages on social media that strives to be independent of the domain to which analysis of messages is applied. However, the limitation here is that words which are not available in the dictionary generated from training dataset are replaced with a generic identifier. This could have an adverse effect on sentiment prediction as there are many relevant words in the tweets that don't have counterparts in the dictionary, for example the names of Apple's devices, URLs of websites that post news about the company, etc.

Stock market predictions have become an interesting research area, correlation of social sentiment data about a company and its stock values there exists research papers that provide solid efficacy to perform a time series analysis on prediction of stock prices and ensemble models that increases the accuracy of the prediction by performing a sentiment analysis on the co-related socio-economic data of that particular company, though its limitation was performed on a 24-hour interval [18–20]. This research paper extends this notion by performing minute-wise stock price sentiment analysis that gives you a more through window for predicting stock rise and stock fall.

Bharathi et al. [13] used a combination of both sensex points and Really Simple Syndication(RSS) feeds for prediction. They extracted headlines from RSS feeds of major news websites and performed sentiment analysis on the text to establish a correlation between stock market values and the sentiments in the headlines. They produced an improvement of 14.43% as compared to standard algorithms like ID3. The proposed system aims to improve on their research by expanding on the methodology in two ways - 1) reduce the gap between consecutive predictions from 5-day averages to per-minute values and 2) improve the scope and quality of the text used for sentiment analysis by considering tweets from people all over the world instead of the news articles published by a few reporters working for media organizations.

3   Proposed Methodology

Twitter was considered to be the source for the dataset because many companies practice public relations via tweets and also as it provides a concrete API with filtering that would prove imperative to the selection criteria for a specified category of text data. Twitter's limitation of 280 characters per tweet also reduces the possibility of verbose text which would prove difficult to classify. The collected twitter dataset will be pre-processed for any missing inconsistencies, and cleaned using our custom data-cleaning libraries. After preprocessing, a subset of that dataset is manually labeled with a sentiment-value. A Random Forest Classifier is used to classify the rest of the sentiment based on the labeled dataset. For stock price prediction, the stock market data was downloaded from Finam and after it undergoes pre-processing, the processed prices dataset along

with the labeled sentiment dataset is run through a LSTM model. A graphical overview of the system structure is shown in Fig. 1.
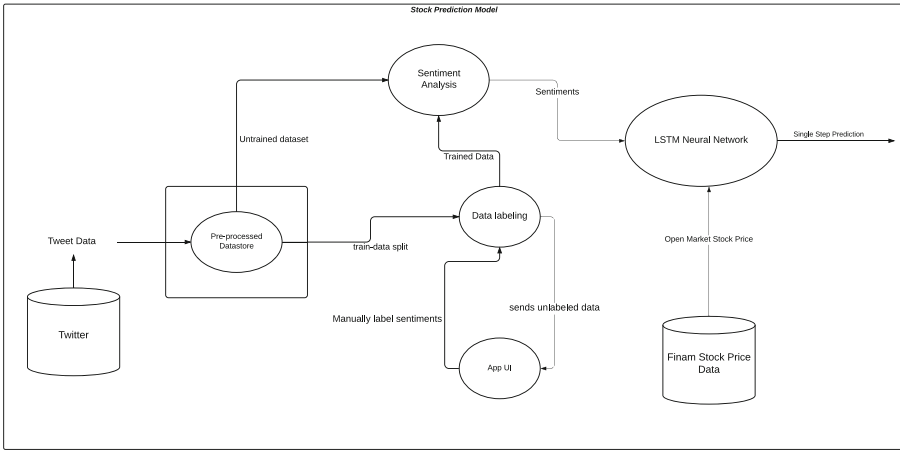

Fig. 1. Proposed system structure

3.1 Data Collection

Data collection is defined as "the process of acquiring raw, unprocessed data and storing in a mutable format". The data collection period was a little over three months, and approximately two million tweets were scraped for the last quarter of 2018. For collection of tweets from Twitter, a python module: 'Twitter-Scraper' was used. It supports querying of the Twitter database with advanced search parameters and operators (Available parameters include followers_count, friends_count, and also the logical operators AND, OR and NOT) [8] that limits its results to tweets that match our query, as well as additional metadata such as ensuring that the time of the tweet fits inside our selected time period. The exact search query given to the module is 'apple OR ((bullish OR bearish) AND (AAPL OR apple))'. This query has proven effective in filtering out the majority of completely unrelated tweets from the result set.

The result object returned by this module is a JSON array of tweet objects, where each tweet is a JSON Object with the following fields: username, user id, html, text, likes, retweets, comments, timestamp, profile-picture, profile display-name, etc. An Example of the raw tweet data is represented in Fig. 2.

The fields user_id, text, and timestamp are extracted from the tweets and other unwanted fields are deleted.

There are several services that provide access to historical intraday stock prices for NASDAQ listed companies like Apple [9]. Finam [10] is a Russian website that provides data for the stock, futures, ETF and Forex markets for

```
{
    "fullname": "BofM_Jeremy",
    "html": "<p class=\"TweetTextSize js-tweet-text tweet-text\" data-aria-label-part=\"0\"
lang=\"en\">Considering moving from non-smart watch to an <strong>Apple</strong> Watch - Series 3
with no cellular in order to save a few bucks.  Am I forgoing anything critical by not getting
4+cellular?</p>",
    "is_retweet": 0,
    "likes": 1,
    "replies": 9,
    "retweet_id": "",
    "retweeter_userid": "",
    "retweeter_username": "",
    "retweets": 0,
    "text": "Considering moving from non-smart watch to an Apple Watch - Series 3 with no
cellular in order to save a few bucks.  Am I forgoing anything critical by not getting
4+cellular?",
    "timestamp": "2018-11-18T23:59:35",
    "timestamp_epochs": 1542585575,
    "tweet_id": "1064307161742934017",
    "tweet_url": "/BofM_Jeremy/status/1064307161742934017",
    "user_id": "4645206635",
    "username": "BofM_Jeremy"
},
```

Fig. 2. Raw tweet data

Table 1. Raw finam stock datatset

DATE	TIME	OPEN	HIGH	LOW	CLOSE	VOL
20180926	100100	161.26	161.3	160.74	160.88	267350
20180926	100200	160.91	160.98	160.52	160.76	163520
20180926	100300	160.67	160.86	160.65	160.84	40670
20180926	100400	160.75	160.88	160.72	160.87	34280
20180926	100500	160.84	160.86	160.64	160.7	61070
20180926	100600	160.65	160.78	160.64	160.67	51490
20180926	100700	160.65	160.68	160.49	160.49	69370
20180926	100800	160.53	160.54	160.37	160.37	48950
20180926	100900	160.48	160.48	160.15	160.15	113360
20180926	101000	160.15	160.45	160.15	160.45	103510

research and analysis purposes. Finam provides data only for certain popular capitalized securities, however for these one can avail several months worth of tick data. A representation of the finam stock dataset is shown in Table 1.

3.2   Data Pre-processing

Data preprocessing is a technique which is used to transform the raw data in a useful and efficient format. In this section unnecessary data or noise is removed from the raw text twitter data. Firstly, the raw text data is converted to lower-case. Secondly, text data which contain words that begin with #(hashtags),

@(user mentions) are simply replaced with the actual word content of the hash-tag and username. Thirdly, long URLs are replaced with just the domain name of the URL. For example, https://techcrunch.com/2019/10/19/the-new-iphone-is-ugly/ is replaced with techcrunch. The identification of these words is implemented through regex matching. Then the special symbols like non-english characters are removed.

The final step in pre-processing of text is stop-word removal, which is the removal of words in the text that do not contribute to the overall meaning of the post. Examples of such words include a, an, the, I, for, etc. The text of each post is tokenized and compared with any publicly available curated list of stop words [11]. The above preprocessing steps were repeated for the remaining two million raw text data. The data preprocessing outcome of an instance is displayed in Table 2.

Table 2. An instance of data pre-processing outcome.

Before data pre-processing	#Apple could have made all their productsćurrent design years ago, now their products are now left behind by @Samsung and others http://gizmo.do/Twg8I8i
After data pre-processing	Apple could made products current design years ago products left behind samsung others gizmodo

3.3   Sentiment Analysis Module

After collecting a large twitter dataset, sentiment analysis is performed on the text. For this purpose, the library Word2vec [12] is used, which is an advanced Natural Language Processing (NLP) technique for mapping words to a vector representation of any dimension. A 200-dimension vector is used for generation in this case. When run on the dataset of text, Word2vec will generate a unique vector for every word in the dataset which will exactly preserve the context of the words and the relation between similar meaning words in vector space. Then the Word2vec representations along with around 15000 messages manually labelled as positive (1), neutral (0) or negative (2). An android app was developed with google's firebase backend systems. The app was distributed to a group of trained people for labeling the tweets. The app contains three buttons for inputting the sentiments below the text data. An instance of the tweet in the app is shown in Fig. 3.
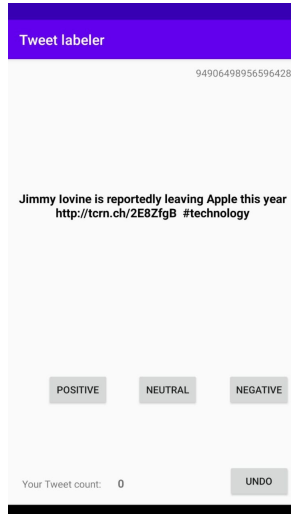
Fig. 3. An instance of app tweet data with sentiment labels.

The output of the app tweet data is then stored within a firebase datastore along with the labeled sentiment. This is highlighted in Fig. 4.

The final labelled dataset had a collection of 7500 neutral tweets, 4201 positive tweets and 3322 negative tweets.

The manually labeled tweets are first split into a training and validation set (by an 80-20% split) to train the random forest classifier. The XGBoost library [14] was used to automatically produce a good set of training parameters for the model. The algorithm was allowed to use 1200 estimators and to reach a maximum depth of 8. On the validation set, the random forest reached a precision score of 90%, Recall of 88% and F1 score 90%. Additionally, K-Fold cross validation was performed on the random forest classifier to ensure the validity of the results. The dataset was split into 10 partitions and each partition was used for validation one at a time, while the other 9 partitions were used to train the classifier. Across the 10 partitions, the classifier reached a mean accuracy of 85.46%.

The trained model is used to predict the sentiments for all the two million tweets in the datastore. To determine the stock price at a future point of time, the random forest classifier predicts social sentiments, and calculates the total sentiment for each one-minute interval as the number of positive minus negative sentiments. Some sample outputs from the random forest classifier are shown in Table 3.

Table 3. Five samples of random forest classifier output.

Text	Timestamp	Sentiment
Liveblog news apple making brooklyn event bit technology news tech	2018-10-27 14:55:00	1
spotify apple music	2018-11-16 19:56:08	0
Ready upgrade tv room ultimate home theater enter usa home entertainment sweepstakes chance win top line samsung 4k uhd tv apple tv 4k bose sound system swee ps	2018-11-08 21:22:05	0
Tech news trending apple launches portal users download data bloomberg uuid 11e8 url	2018-10-17 18:14:20	1
Depends type apple watch want come different sizes different bands cost	2018-12-25 20:40:32	1

Fig. 4. An instance of labeled tweet data.

4   Results and Discussion

A time series forecasting method using LSTM is used here as both the social media posts as well as the financial stock price dataset has a time component. LSTM was trained using Google's open source TensorFlow libraries which comes with an implementation of LSTM. The model contained one LSTM layer with 32 units, densely connected to a second layer with one neuron that used the sigmoid activation function. The RMSProp algorithm [15] was used to speed up the training process.

As the dataset contains two distinct features namely the sentiment and the stock price, a multivariate version of the dataset was created for training. In this process, overlapping sliding windows of length 720 min (12 h) are applied on the dataset. Within the window the value is sampled every 10 min in order to smooth out the smaller variations every minute while still retaining the rises and falls of higher magnitude. Hence each window contains 72 points of data. To get a single step dataset, the starting point of the window is set to the point immediately after the start of the previous window, i.e. a new window begins every 10 min. Some samples of the data in each window are shown in Figs. 6, 7 and 8.

Finally the dataset contains 35634 entries of data, which is divided into a testing and validation set (80-20 % ratio). The LSTM was trained on the dataset for a total of 50 epochs. The training and validation loss for a dataset consisting of windows of history size 720 min, and step size 10 min is visualised in Fig. 5.
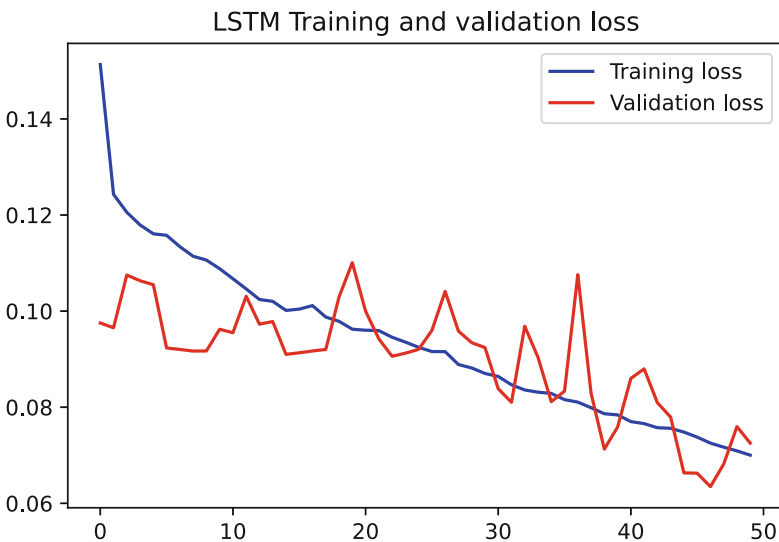

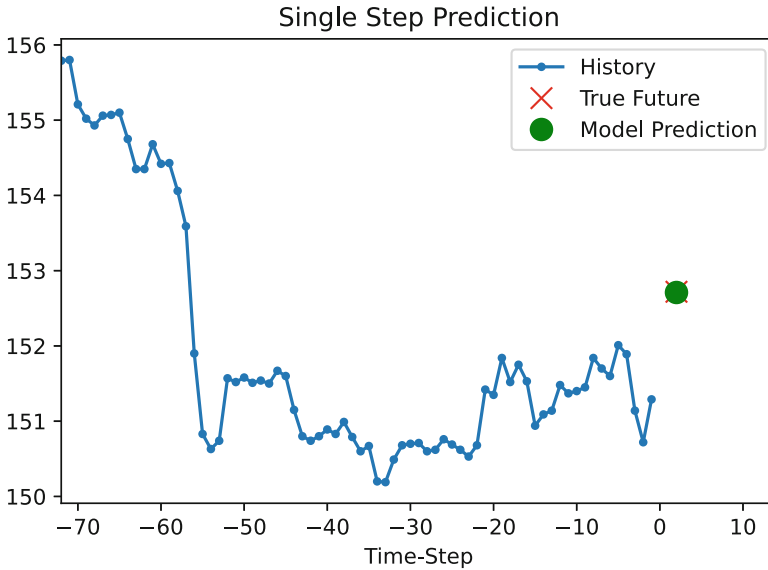**Fig. 5.** LSTM training and Validation loss

Fig. 6. First sample prediction of LSTM (Color figure online)

The first prediction of the trained LSTM model at a future time from the validation set is shown in Fig. 6. The blue line shows a subset of the stock price history passed as input to the LSTM. The green circle represents the price predicted by the model (152.71), whereas the red cross shows the actual value that came to pass at that point of time (152.72). From Fig. 6, it is observed that the model is accurately predicting the sentiment of the tweet.

The second prediction of the trained LSTM model at a future time from the validation set is shown in Fig. 7. The predicted price was 148.28 whereas the observed actual price was 148.49. From Fig. 7, it is observed that the model is accurately predicting the sentiment of the tweet.

The third prediction of the trained LSTM model is shown in Fig. 8. The predicted price was 152.42 whereas the observed actual price was 151.64. From Fig. 8, it is observed that there is a deviation between the model prediction and the actual value. Here an instance of inconsistency with the prediction is observed. Though there are minor deviations in some single step predictions, this process usually consists of a large number of single step predictions made over a period of time which allows the algorithm to make an accurate prediction overall.

While the discussion above shows the results of training with windows of history size 720 and step size 10 min, experiments were also conducted with windows of varying parameters. The Mean Squared Error(MSE) of predictions on the validation set using these parameters was calculated.
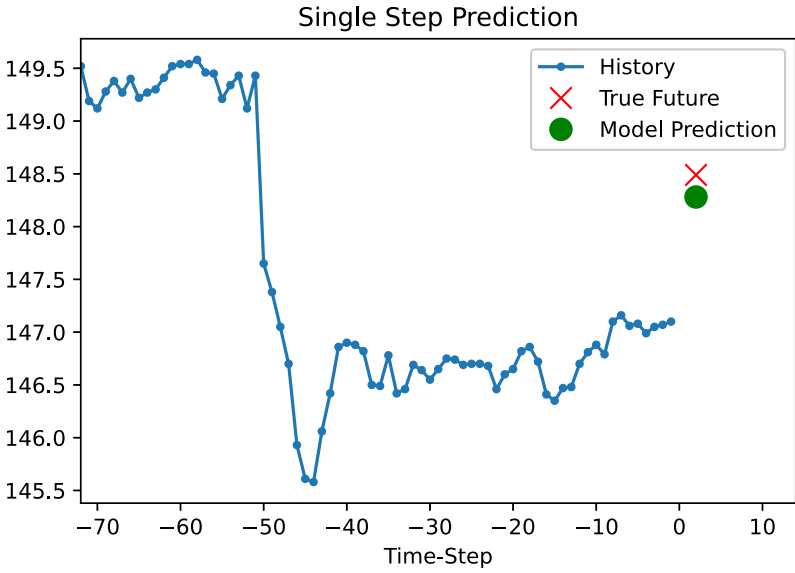
Fig. 7. Second sample prediction of LSTM (Color figure online)
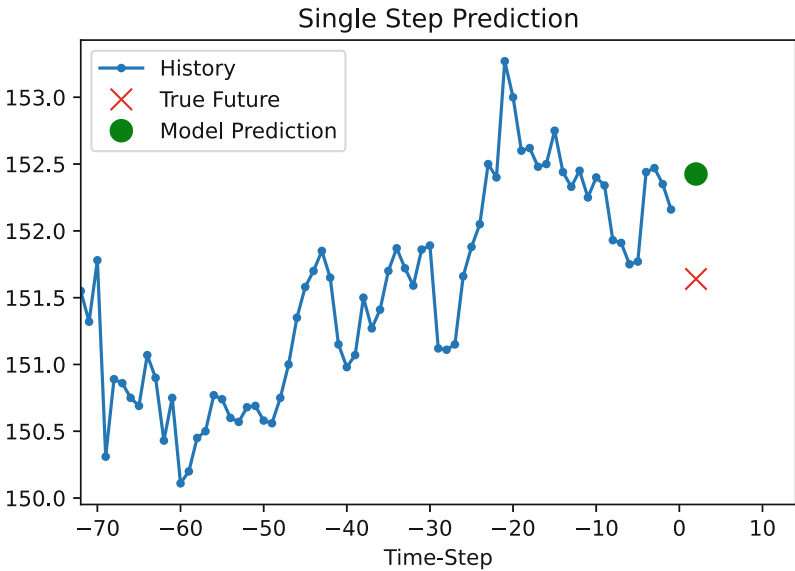

Fig. 8. Third sample prediction of LSTM (Color figure online)

Additionally, in order to support the claim that combining sentiment values with the pricing dataset produces a better accuracy, the process of training the LSTM was repeated after dropping the sentiment information from the dataset, keeping the other parameters constant. The MSE of predictions without sentiment data was also calculated for various window sizes and is shown in Table 4. It is to be noted that the value of MSE without sentiment is consistently higher than the corresponding value for the dataset with sentiment, thus supporting the claim.

Table 4. Variation of MSE with different History and step size for datasets with and without sentiment data.

History size	Step size	MSE	MSE without sentiment
720	50	38.47090527	67.73315905
720	10	35.05139267	52.17865264
720	80	28.47415585	36.00919354
360	50	36.9319954	35.84480859
60	20	35.79312783	34.83763717
30	10	32.89082675	33.12358949
15	5	5.930060326	34.24569216
15	10	5.187107316	14.64130006

Further, it is observed from Table 4 that as the step size and history size decreases, the MSE also reduces, indicating an improvement in the accuracy rate.

Table 5. Stock values prediction and errors.

| Date	Time	Actual price	Predicted	Error
20181218	155200	151.42	151.68074	−0.26074
20181218	155300	151.27	151.69843	−0.42843
20181218	155400	151.44	151.9058	−0.4658
20181218	155500	151.49	151.82811	−0.33811
20181218	155600	151.27	151.72792	−0.45792
20181218	155700	151.33	151.73984	−0.40894
20181218	155800	151.35	151.96754	−0.61754
20181218	155900	151.34	151.75021	−0.41021
20181218	160000	151.36	151.98468	−0.62468
20181218	160100	151.35	151.85138	−0.50138
20181218	160200	151.35	151.78558	−0.43558
20181218	160300	151.46	151.78085	−0.32085
20181218	160400	151.55	151.88606	−0.33606
20181218	160500	151.61	151.85497	−0.24497
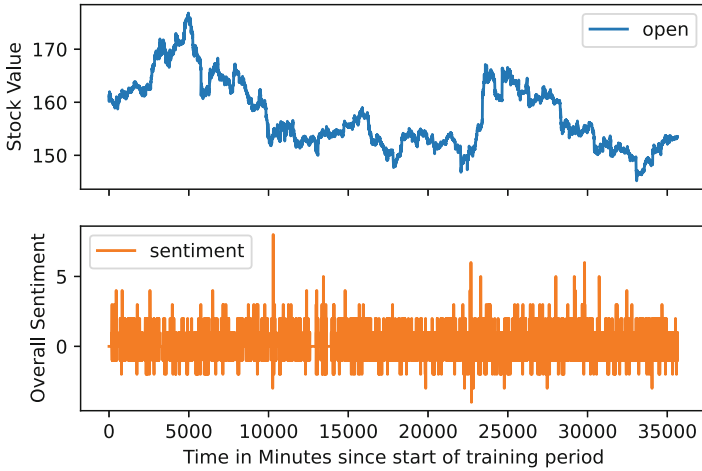
Fig. 9. Visualization of processed training dataset (Color figure online)

Table 5 displays the ground truth and the predicted price for a portion of the validation set. It is observed that the LSTM is able to learn and accurately predict with errors less than 1 in the vast majority of cases at each minute interval. The system is well trained over various curves, rises, falls over the last quarter of 2018.

Though there are a multitude of factors which could determine the exact values of stock prices, based on the outcome of the system. The public image of the company seems to be one of the driving forces. The LSTM Neural Network was accurate enough to forecast the stock values, which indicates the existence of a correlation between the sentiments and stock prices. This suggests that the public image of a company has a bearing on the market performance of a company. The visualisation found between the stock prices and sentiments are shown in the Fig. 9. The y-axis for the blue line shows the market price and the total sentiment is shown in the orange graph. The x-axis represents the time component, i.e. the number of minutes since the start of Q4 2018.

5   Conclusion and Future Work

In this paper, the rise and fall of stock prices were predicted at every minute interval. Intraday (1 min intervals) stock market data on Apple Inc. (NASDAQ: AAPL) was collected for Q4, 2018. Twitter was scraped to find all tweets related to Apple over the same time period. Total sentiment in each one-minute interval was calculated and combined with market price histories to forecast the future prices. Confirming the hypothesis there exists a correlation between a company's stock price and its public perception.

Posts on social media usually include one or more emoticons. These are more capable of conveying what people are feeling than plain text data. They are espe-

cially useful in identifying sarcasm in messages which may reduce false positive or negative classification of sentiments. Another aspect of data collection to be explored is to rank tweets based on the number of retweets and likes. This would add a biasing factor which would determine the stock prices more accurately. Government Authorities and Diplomats would have more say in impacting the stock market as a whole.

Once the expansion of the prediction module is completed with favourable results, the authors are interested to deploy the implementation as part of a program capable of automatically buying and selling shares of the company, based on real time feedback from the market and social media. Since there is already research being done on this aspect by using stream processing softwares as described by Behera, R. K., et al. [17], the authors hope to build on the progress already made and implement the model, so that the project may have a practical benefit aside from the research arena.

References

1. Mankar, T., Hotchandani, T., Madhwani, M., Chidrawar, A., Lifna, C.S.: Stock market prediction based on social sentiments using machine learning, pp. 1–3 (2018) https://doi.org/10.1109/ICSCET.2018.8537242
2. Acosta, J., Lamaute, N., Luo, M., Finkelstein, E., Andreea, C.: Sentiment analysis of twitter messages using Word2Vec. In: Proceedings of Student-Faculty Research Day, CSIS, Pace University (May 5th 2017)
3. Venkata, S.P., Kamal, N.C., Ganapati, P., Babita, M.: Sentiment analysis of twitter data for predicting stock market movements. In: International conference on Signal Processing, Communication, Power and Embedded System, pp. 1345–1350 (2016)
4. Rakhi, B., Sher, M.D.: Integrating StockTwits with sentiment analysis for better prediction of stock price movement. In: IEEE International Conference on Computing, Mathematics and Engineering Technologies – iCoMET (2018)
5. Coyne, S., Madiraju, P., Coelho, J.: Forecasting stock prices using social media analysis. In: IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 15th International Conference on Pervasive Intelligence and Computing, 3rd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress, pp. 1031–1038 (2017)
6. Wang, Z., Ho, S.B., Lin, Z.: Stock market prediction by incorporating social media news as sentiment https://ieeexplore.ieee.org/document/8637365
7. Selvin, S., Vinayakumar, R., Gopalakrishnan, E.A., Menon, V.K., Soman, K.P.: Stock price prediction using LSTM. RNN and CNN-sliding window model, pp. 1643–1647 (2017). https://doi.org/10.1109/ICACCI.2017.8126078
8. Standard search operators available in the Twitter search query field: https://developer.twitter.com/en/docs/tweets/rules-and-filtering/overview/standard-operators
9. Publicly Available sources o Intra-day stock market data for listed companies https://www.quantshare.com/sa-636-6-new-ways-to-download-free-intraday-data-for-the-us-stock-market
10. Finam.ru - A website that provides several months of tick data for highly capitalized securities https://www.finam.ru/profile/moex-akcii/gazprom/export/

11. Curated list of English stop-words extracted from Python's NLTK library: https://gist.github.com/sebleier/554280
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space, pp. 1–12 (2013)
13. Bharathi, S., Geetha, A.: Sentiment analysis for effective stock market prediction. Int. J. Intell. Eng. Syst. **10**, 146–154 (2017) https://doi.org/10.22266/ijies2017.0630.16
14. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. New York, ACM (2016) https://doi.org/10.1145/2939672.2939785
15. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. COURSERA: Neural Netw. Mach. Learn. **4**(2), 26–31 (2012)
16. Behera R.K., et al.: Co-LSTM: convolutional LSTM model for sentiment analysis in social big data. Inf. Proc. Manage. **58**(1), 102435 (2021)
17. Behera, R.K., et al.: Comparative study of real time machine learning models for stock prediction through streaming data. J. Univ. Comput. Sci. **26**(9), 1128–1147 (2020)
18. Valle-Cruz, D., Fernandez-Cortez, V., López-Chau, A., Sandoval-Almazán, R.: Does Twitter affect stock market decisions? financial sentiment analysis during pandemics: a comparative study of the H1N1 and the COVID-19 periods. Cogn. Comput. 1–16 (2021). https://doi.org/10.1007/s12559-021-09819-8
19. Nti, I.K., Adekoya, A.F., Weyori, B.A.: Predicting stock market price movement using sentiment analysis: evidence from ghana. Appl. Comput. Syst. **25**(1), 33–42 (2020). https://doi.org/10.2478/acss-2020-0004
20. Carosia, A.E.O., Coelho, G.P., Silva, A.E.A.: Analyzing the Brazilian financial market through Portuguese sentiment analysis in social media. Appl. Artif. Intell. **34**(1), 1–19 (2020)